Deep Neural Networks for Learning Spatio-Temporal Features From Tomography Sensors

Omar Costilla-Reyes, Patricia Scully, and Krikor B. Ozanyan, *Senior Member, IEEE*

***Abstract*—We demonstrate accurate spatio-temporal gait data classification from raw tomography sensor data without the need to reconstruct images. This is based on a simple yet efficient machine learning methodology based on a convolutional neural network architecture for learning spatio-temporal features, automatically end-to-end from raw sensor data. In a case study on a floor pressure tomography sensor, experimental results show an effective gait pattern classification *F*-score performance of 97.88 $\pm$ 1.70%. It is shown that the automatic extraction of classification features from raw data leads to a substantially better performance, compared to features derived by shallow machine learning models that use the reconstructed images as input, implying that for the purpose of automatic decision-making it is possible to eliminate the image reconstruction step. This approach is portable across a range of industrial tasks that involve tomography sensors. The proposed learning architecture is computationally efficient, has a low number of parameters and is able to achieve reliable classification *F*-score performance from a limited set of experimental samples. We also introduce a floor sensor dataset of 892 samples, encompassing experiments of 10 manners of walking and 3 cognitive-oriented tasks to yield a total of 13 types of gait patterns.**

***Index Terms*—Convolutional neural networks (CNNs), deep learning, floor sensor system, machine learning, spatio-temporal analysis, tomography.**

Manuscript received March 8, 2017; revised May 15, 2017; accepted June 3, 2017. Date of publication August 9, 2017; date of current version November 16, 2017. This work was supported by the U.K. Engineering and Physical Sciences Research Council EP/K005294/1 EP/K503447/1 and in part by the Consejo Nacional de Ciencia y Tecnologia under Grant 467373. *(Corresponding author: Omar Costilla-Reyes.)*

O. Costilla-Reyes and K. B Ozanyan are with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: omar.costilla.reyes@gmail.com; k.ozanyan@manchester.ac.uk).

P. Scully is with the School of Chemical Engineering and Analytical Science, Faculty of Engineering and Physical Sciences and the Photon Science Institute, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: patricia.scully@manchester.ac.uk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material contains a 3D gait surface visualization provided (as a MATLAB figure, to allow better inspection) together with a gait progression movie of the normal walk experiment using the spatio-temporal reconstruction approach introduced in the paper "Deep Neural Networks for Learning Spatio-Temporal Features from Tomography Sensors." The total size of the file is 18 MB. Contact omar.costilla.reyes@gmail.com for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2017.2716907

I. Introduction

WITH the industrial world moving further into numerous variants of smart sensing, the already established area of industrial imaging needs to reassess long-standing paradigms in the light of the new opportunities and challenges. Indirect imaging, such as tomography, has played hitherto the role of an important utility because of problems with direct access to industrial subjects: spatial limitations introduced by physical restriction, as well as temporal limitations caused by requirements for speed and volume of data [1]. Furthermore, tomography is a compressed sensing method [2] and takes direct advantage of sensor fusion [3], which renders it quite efficient in the context of smart sensing. Tomography sensors deliver measurement data that are a spatio-temporal sample of the imaged object. A popular approach is to treat the spatial aspect by an inverse problem solution while seeking time correlation between images reconstructed at known times.

In the example case of industrial rheology, time sequences of reconstructed cross sections of flow have been used to present three-dimensional (3-D) models [4], [5]. Beyond this substantial achievement toward flow visualization, the obtained cross-sectional images and 3-D presentations still need interpretation, in terms of various flow regimes and their transitions, required for controlling an industrial process. This has hindered progress in the direct use of tomography for automatic process control since sophisticated image analyses by humans are needed in order to extract input parameters. The problem is exacerbated by the futility of defining meaningful flags based on a small number of measures and the lack of portability of such an approach across variations in the process conditions.

This paper addresses the problem by removing the image reconstruction stage and, thus, the need for image interpretation. Instead, the information contents of the spatio-temporal raw data acquired from tomography are used directly for machine learning to enable classifications of the process conditions based on quantities suitable as inputs for process control. Initial work on software and hardware elements of machine learning from tomography data, to estimate the component fractions of multiphase flows, as well as the heights and orientation of the components interfaces has been reviewed in [6]. It is worth noting that in these early reports no spatio-temporal fusion has been attempted and classifications were based on features that would be straightforward to extract from cross-sectional images (areas,

distances, the orientation of lines) although the latter were not explicitly reconstructed. In this paper, we introduce a machine-learning model based on a convolutional neural network (CNN), a form of deep learning [7], for pattern classification and a raw sensor data transformation technique that allows the automatic extraction of features from the raw spatio-temporal tomography sensor data. Since CNN models are usually applied for direct image processing problems [7] (e.g., images from cameras), the proposed technique considers the measured spatio-temporal data as a direct sensor amplitude matrix that is used on a CNN model. Experiments with costly tomography sensors on an industrial plant are difficult to conduct and may be impractical in view of the need for a scale-down, therefore, we pilot our approach on tomography data acquired with a floor pressure imaging sensor (iMAGiMAT) for the classification of manners of walking. Analogous to extracting features from flow cross-sectional images, here gait features can be extracted from reconstructed footstep images [8].

Similar to flows and other industrial subjects, gait can be analyzed visually by humans or automatically from suitable sensor data. Sensor systems for automatic gait analysis [9] have relied heavily on man-made features, such as the body's centre of pressure, stride length, and cadence, influenced by the limitations of the existing practice of observation. Thus, the crux of the problem, the dependence of analysis on human perception, as well as the uncertainty in the correct choice of classification features are common in a wide range of industrial scenarios. Consequently, seemingly far apart examples, such as the characterizations of flow and gait, are amenable to a common solution for the automatic extraction of classification features from complex spatio-temporal process data.

In a case study of the performance of this method, we introduce and use a gait dataset, UoM-Gait-13, which contains floor pressure signals recorded from 13 experiments: 10 different manners of walking and 3 dual-task experiments: walking while performing another cognitive task [10]. The proposed approach yields a top classification F-score performance of 97.88 $\pm$ 1.70% to identify the 13 types of gait patterns in the dataset. The proposed CNN architecture is also computationally efficient since only a small number of CNN weights needs to be trained. As opposed to other CNN models that require large datasets for training [7], our CNN model was able to learn a robust set of representations from a limited dataset. The rest of the paper is organized as follows: In Section II, the background is presented, in Section III, the spatio-temporal raw data are explained, and in Section IV, the collected gait dataset is described. The methods and procedures are presented in Section V and experiments are presented in Section VI. The results are presented in Section VII. Finally, the discussion and conclusions are presented in Section VIII.

II. Background: Floor Sensors, Gait Analysis, and Machine Learning

A. Sensing for Gait Analysis

Gait analysis has been widely studied for a variety of applications including healthcare, biometrics, sports, etc., [11]. Robot gait (biped, quadruped, and hexapod robots) is one notable gait application relevant to industrial electronics [12]. Classification of a person's given its emotional state has also been explored. A person's pride, happiness, neutral emotion, fear, and anger has been classified with high statistical confidence given only its gait pattern [13]. Generally, three types of gait monitoring systems exist, namely: cameras using image processing, floor sensors, and wearable sensors [14]. The use of cameras for gait is vulnerable to details in the environment such as lighting and requires walking in a straight line, perpendicular to the position of the camera to capture gait images optimal for analysis. Besides that the use of cameras is considered as invasion of privacy in living environments, e.g., for healthcare [15], because of disadvantageous parallels to video surveillance. The disadvantage of wearable sensors is that the sensors need to be attached to the body, may be uncomfortable to wear, as well as require assistance to attach correctly. On the other hand, floor sensor systems have the advantage of being noninvasive and even unobtrusive, less prone to environmental noise, and undemanding the subject's attention, which affects the data quality positively.

B. Floor Sensors for Gait Analysis

Only a few studies have focused on detecting gait activities with a floor sensor system similar to the experiments presented here. Saripalle *et al.* [9] classified choreographed movements from subjects whilst standing. The classification scores range from 60% to 100% depending on the features, type of movements and classifiers used. The features given to classifiers were based on the careful selection of coefficient of performance parameters. Headon and Curwen [16] used a hidden Markov model for classification of movements based on the values of the footstep ground reaction force, the classification accuracy is close to 100%. These previous studies rely heavily on hand-made feature engineering to classify gait activities. The advantage of the approach presented in this paper is that it learns a robust set of representations from spatio-temporal raw sensor data.

C. iMAGiMAT Floor Sensor System

iMAGiMAT, an experimental floor sensor system demonstrator for gait analysis [8], was employed to acquire the UoM-Gait-13 dataset used for the analysis in this paper. Deformation of plastic optical fibres (POFs) due to applied footstep pressure, modulates their transmission properties and, therefore, the intensity of the transmitted light. The system comprises 116 POFs distributed sensors strategically placed at 3 angles as shown in Fig. 1. The manufacturing cost of the floor sensor system has been estimated at under 150 USD per square meter [8], which is more than an order of magnitude less than commercial alternatives (e.g., the GAITRite walkway)

The extremely low number of angular projections and the total number of measurements result in a severely under-sampled spatio-temporal distribution of the floor deformation field. However, this is typical for industrial tomography [1], because of the access constraints as already mentioned in Section I, additionally justifying the use of the sensor and database in this paper. Using general tomography theory and/or special limited-view tomography approaches, footstep reconstructions are possible, e.g., by the Landweber image reconstruction algorithm [17],

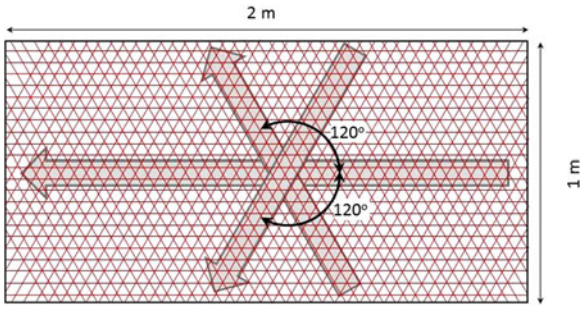

Fig. 1. Conceptual drawing of the sensor head of the iMAGiMAT floor sensor system. The red lines across the sensor head area (1 × 2 sq.m.) represent the optical fiber attenuation line integrals, grouped in three angular projections to ensure consistent spatial and temporal sampling.

to be addressed in Section V-B, demonstrating that the dataset contains adequate spatial information at the frame rate applied.

The spatial component of the floor sensor data is described by signals obtained at any moment from a set of 116 POF sensors at 3 different angles, as presented in Fig. 1. From a tomography point of view, this is equivalent to taking a Radon transform from three angular projections at 120° [18]. The floor sensor geometry allows capturing the spatial distribution of pressure, while the time component is captured as frames acquired at 256 Hz, thus, providing the time evolution of the spatial component.

D. CNNs to Learn Spatio-Temporal Features

Deep learning models deliver state-of-the-art performance for image recognition, speech recognition, and other applications [7]. One of the most studied spatio-temporal problems is to recognize human actions from videos [19]. In recent years, the top performing models to solve this problem have been based on CNNs and recurrent neural networks [19]–[21]. The proposed architectures use large publicly available video datasets. These approaches are effective to learn representations from raw video frames. The disadvantage is that they are complex and require large computational resources to train them; furthermore, in some cases, other preprocessing steps are required, such as to calculate optical flow between video frames [19].

III. Visualization of the Raw Spatio-Temporal Sensor Data

Fig. 2 shows the floor sensor system raw signals in the time domain for four gait experiments. Fig. 2(a) shows the amplitude response over time of the 116 spatial sensor signals of 1 sample of the normal gait experiment. Each signal has a different amplitude response that is determined by the physical characteristics of each POF sensor and the executed gait pattern type. The sensors that show a constant signal amplitude over time are not active for that sample. The spatial average $\mathrm{SA}[t]$ function is defined by

$$\mathrm{SA}[t] = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{SF}_i[t]) \tag{1}$$

where t represents the time component in frames. SF_i is the amplitude response of the ith POF sensor of the floor sensor system and n is the number of POF sensors. In Fig. 2(b), (c), and (d) the $\mathrm{SA}[t]$ function can be observed for the experiments of normal gait, slow gait, and barefoot gait, respectively. A total 20 samples per experiment are shown in each figure. The $\mathrm{SA}[t]$ function allows observation of the gait patterns in the time domain. The variation in the amplitude responses of experiments belonging to the same class is due to the user stepping into different sensor area within the floor sensor system, and the time delay between signals is due to different gait execution time in experiments.

In this paper, the spatial and temporal components of the dataset are considered as they are without further processing stages as alignment or rotation. Fig. 2 suggests that the pattern recognition problem addressed here is not trivial, since the patterns between gait experiments are similar, thus, not easy to differentiate, at least not for the human eye. The $\mathrm{SA}[t]$ transformation is used as a temporal feature for the analysis presented in [22].

IV. UoM-Gait-13: Spatio-Temporal Dataset

The dataset experiments were divided into ten manners of walking and three dual tasks. The dataset consists of a total of 892 samples with an unbalanced number of samples per class. We expanded the dataset presented in [22], where a top cross-validated F-score classification performance of $90.84 \pm 2.46\%$ was reported by using temporal domain features and a Random Forest classifier for ten gait classes. The dataset presented here adds a further three dual-task experiments [10] to the existing dataset: reading, naming animals, and counting backward while walking normally. The inclusion of the dual-task experiments aims to increase the complexity of the task and establish the detectability of the perturbation on the gait patterns since the change in gait is less pronounced for a double-task than between manners of walking.

The gait experiments were performed along the 2-m length of the floor sensor system, which allowed a longer gait to be captured. We followed the methodology presented in [10] for the gait experiments.

The dataset characteristics are presented in Table I. Each sample contains 116 sensors signals obtained in a time window of 5.4 s (1400 frames at 256 Hz). The time window was selected based on the average time that the experiments lasted. Therefore, the signals were repeated or truncated in time to form a standard length of 1400 frames (see Fig. 2). The dataset is available to the research community by request to the main author.

V. Methods for Spatio-Temporal Analysis

A. Spatio-Temporal Raw Sensor Matrices (RSMs)

1) *Feature Selection:* Our spatio-temporal raw sensor technique considers the floor raw sensor data as a sensor matrix at each frame step to create a spatio-temporal volume, which is then considered as input to a CNN model. As a first step, a linear support vector machine (SVM) classifier was trained per POF sensor signal (116 models in total) of each sample in the dataset for sensor feature selection. A dataset split of 50% for training and 50% for testing was applied from the 892 samples available. The large 50% testing set split aims to improve the

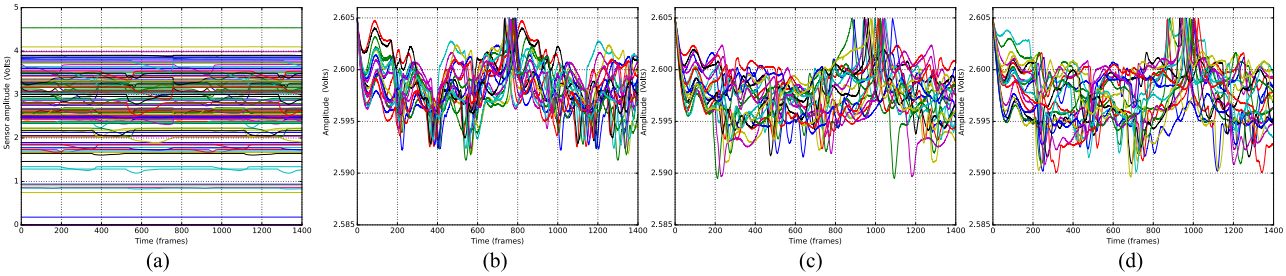

Fig. 2. Raw sensor signals for the normal walking experiment. (a) Plot of each POF amplitude (volts) over time (frames); (b), (c), and (d) SA feature response for 20 samples of class 1, 2, and 4, respectively.

TABLE I
UOM-GAIT-13 DATASET DESCRIPTION

Class	Experiment	Experiment Description	Steps	Samples
1	Normal gait	Walking at normal speed and gait frequency (cadence).	4	92
2	Slow gait	Lower constant gait speed and cadence compared to class 1.	4	84
3	Fast gait	Higher constant gait speed and cadence compared to class 1.	3	113
4	Barefoot gait	Normal gait, but with modified GRF distribution.	4	88
5	Gait with weight	Normal gait attempted while carrying a 10 kg bag with both hands.	4	88
6	Gait with hands back	Normal gait attempted with hands crossed behind the back.	4	82
7	Backward gait	Gait with modified walking pattern, cadence, stride and centre of pressure locus	4 or 5	81
8	Right foot leading gait	Gait progression one step at a time, with the right leg leading.	4 or 5	54
9	Left foot leading gait	Gait progression one step at a time, with the left leg leading.	4 or 5	46
10	Side gait	Walking sideways, with the coronal plane aligned with the direction of movement (either leg leading).	4 or 5	42
11	Naming animals task	Walking at normal speed while pronouncing animals'names out loud.	4 or 5	49
12	Counting backward task	Walking at normal speed while performing serial seven subtractions from a random three digit number.	4 or 5	27
13	Reading task	Walking at normal speed while reading a book.	4 or 5	46

generalization performance of our model to unseen data. It also indicates that the model is able to generalize optimally since it was only trained with a 50% split of the dataset. This experiment allowed us to discard six POF sensor signals with the lowest classification performance to fit the sensor data into a rectangular matrix; therefore, only 110 POF sensor signals were considered for our spatio-temporal raw sensor approach The classification performance of the six discarded POF is lower than 50% *F*-score.

2) *Representation:* The 110 sensor signals, available at each frame of the samples in the dataset, were placed in a matrix of 11 (height) by 10 (width) spatial dimensions. The first row of the matrix contains the amplitude response of the first 10 set of sensors, the second row the second set of 10 sensors, etc., for the 110 sensors considered. In this spatio-temporal RSM, each sensor signal is represented as a pixel-valued colour map according to the signal level obtained at the nth frame as shown in Fig. 3 for a frame set.

2) *Dimensionality Reduction and Visualization:* The sensor signals were also downsampled in the time domain, with a down-sampling factor of five found to be optimal, by using an order 8 Chebyshev type I filter [23]. The resultant time dimension obtained was 280 frames, reduced from the original 1400 frames available per sample. The procedure improved the classification accuracy and reduced computation time compared to using the dataset without the down-sampling procedure. Fig. 3 shows the spatio-temporal RSM for the following down-sampled frames: 1, 15, 20, 25, and 30 of one sample of the normal walk experiment, the sequence was selected to allow comparison with the trained feature maps presented in Section VII-E. The spatio-temporal RSMs volumes obtained from the UoM-Gait-13 dataset were used as input volumes to the proposed CNN model.

B. Tomography Spatial Reconstruction of Foot Pressure

Spatial time-integrated gait experiments reconstruction were performed with the aim to allow comparison between the CNN performance in the two cases 1) classifications from raw spatio-temporal data with automatic feature extraction and 2) classifications from pixel features generated from reconstructed images.

The Landweber's method is an iterative image reconstruction algorithm, based on finding a minimum of the inverse transform error [17]. This is usually applied in nonmedical tomography applications where the data are typically under-sampled. If n and m denote the number of rows and columns of the data, respectively, then a measurement vector $P(1,n)$ is represented as the product of the transformation matrix $S(n,m)$ with the true pixel vector $F(m,1)$, or $P = S * F$. The inverse problem can be solved using the inverse S^{-1} since $F = S^{-1} * P$. In the Landweber method, the inverse S^{-1} is approximated by the transpose S^{-T} since S^{-1} does not exist (S is rectangular, $m > n$). Thus, the basic mathematical expression used for Landweber iterative reconstruction is

$$F_{n+1} = F_n + \alpha S^T (P - \mathrm{SF}_n) \tag{2}$$

where F_n is the pixel vector representing the reconstructed image obtained at the nth iteration. The empirically chosen

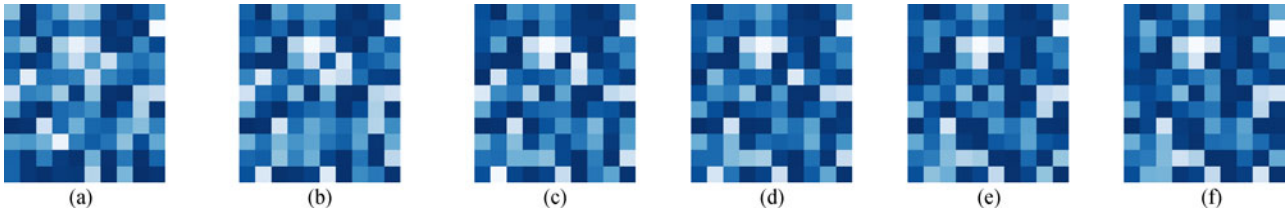

Fig. 3. Set of frames for one sample of the normal walk experiment. White represents a zero output POF sensor value. Light blue: low output POF sensor value, solid blue: high output sensor value. (a) Frame 1. (b) Frame 10. (c) Frame 15. (d) Frame 20. (e) Frame 25. (f) Frame 30.

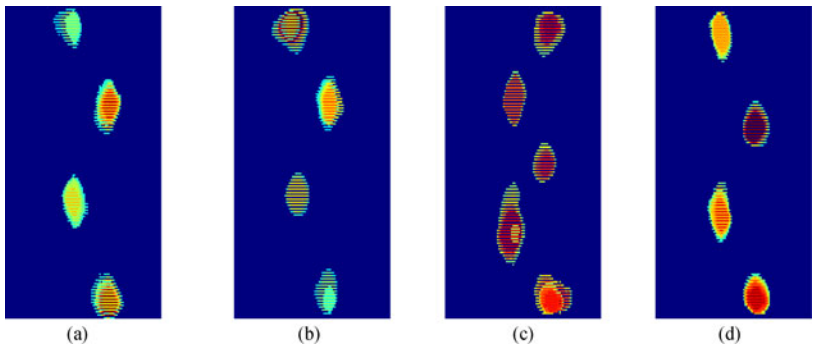

Fig. 4. Spatial reconstruction for one sample of classes 1, 4, 5, 11. (a) Class 1. (b) Class 4. (c) Class 7. (d) Class 11.

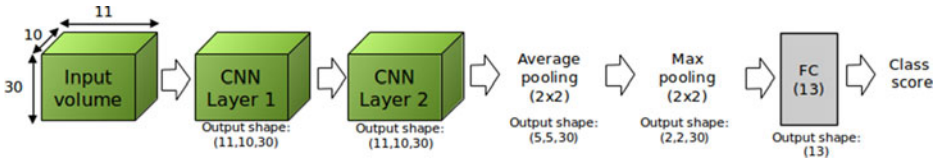

Fig. 5. CNN architecture.

weight α (relaxation factor) affects strongly the algorithm's convergence. Since large values of α obstruct convergence, the recommended upper limit for the relaxation factor is $\alpha < \lambda_{\max}^{-1}$, where $\lambda_{\max}^{-1}$ is the largest eigenvalue of the transformation matrix S and is independent of the measurement vector P.

Fig. 4 shows the spatio-temporal reconstruction of one sample of the following experiments: normal gait (class 1), barefoot gait (class 4), gait with weight (class 5), and the naming animal's task (class 11) by using the Landweber algorithm. The reconstructions are performed on a 200×100 pixel grid, thus, the four subfigures in Fig. 4 are exact spatial maps of the four gait experiments over the complete temporal component, where frame sequences are incorporated by a pixel-by-pixel overlay of the individual frames, with a single pixel corresponding to 1 cm^2. A 3-D gait surface visualization is provided (as a MATLAB figure, to allow better inspection) in the Supplementary media material of this paper, together with a gait progression movie of the normal walk experiment using the spatial reconstruction.

VI. Experiments

A. Setup

The experiments were run on an Ubuntu 14.04 LTS machine. The CNNs models were built with the open source Theano [24] mathematical framework. A NVIDIA Titan X Graphical Processing Unit (GPU) was used to accelerate the model's computation time required for training and testing.

B. CNN Model Architecture for Spatio-Temporal RSMs

Our CNN architecture consists of two convolutional (conv) layers and one fully connected (FC) layer. The first conv layer takes as input a spatio-temporal RSM volume of n frames of 11 (height) by 10 (width) spatial dimensions. The two conv layers have n channels each that assign one frame of the input to a single conv layer channel. The conv layers use a stride of 1 and zero-padding that forces same input to output shape at the conv layers. A filter of 5×5 dimensions was used. The latter use batch normalization and ReLU activation operations as in [25]. The second conv layer is followed by a max-pooling operation to down-sample the spatial component of the volume by a factor of 2. This is followed by an average pooling operation of a factor of 2 for both vertical and horizontal spatial components. The max-pooling operation has the purpose of replacing the last FC layers used in other CNNs architectures [7], in order to reduce the model's complexity, computation time, and the number of parameters. This approach is inspired by the ResNet [25] architecture, which achieved state-of-the-art classification performance in the ImagNet challenge of 2015 [26]. The model architecture can be visualized in Fig. 5. The CNN model architecture presented here has a total of 46,673 parameters. The parameters per layer are 45,060 for the two conv layers, 40 for

TABLE II
CNN MODEL CLASSIFICATION PERFORMANCE FOR VARIOUS SETS OF CONSECUTIVE FRAMES

Frames	Accuracy	Precison	Recall	*F*-score	Training Time (s)	Training Epochs	Epoch Exec. Time (s)
1	11.00 ± 2.30%	4.92 ± 2.02%	11.00 ± 2.30%	5.64 ± 1.68%	87.09	119	0.73
5	96.03 ± 2.39%	96.35 ± 2.20%	96.03 ± 2.39%	96.04 ± 2.39%	81.45	111	0.73
10	96.51 ± 2.10%	97.03 ± 1.80%	96.56 ± 2.06%	96.32 ± 3.40%	83.98	110	0.76
20	96.86 ± 1.90%	97.29 ± 1.70%	96.86 ± 1.90%	96.79 ± 1.96%	83.91	107	0.78
30	**97.88 ± 1.70%**	**98.09 ± 1.50%**	**97.88 ± 1.70%**	**97.88 ± 1.70%**	**51.68**	**66**	**0.78**
50	96.85 ± 2.07%	97.21 ± 1.82%	96.85 ± 2.07%	96.81 ± 2.10%	45.65	57	0.80
100	96.27 ± 2.04%	96.73 ± 1.82%	96.27 ± 2.04%	96.27 ± 2.04%	34.66	39	0.89
150	96.28 ± 2.07%	96.89 ± 1.71%	96.28 ± 2.07%	96.28 ± 2.06%	47.50	44	1.08
200	96.62 ± 1.87%	96.88 ± 1.71%	96.62 ± 1.86%	96.61 ± 1.86%	56.86	44	1.29
280	96.62 ± 1.88%	96.94 ± 1.74%	96.62 ± 1.88%	96.62 ± 1.88%	92.12	48	1.92

the two batch norm layers, and 1,573 for the FC layer. This is a very small number of parameters, when compared to other larger CNNs, such as ResNet.

C. Model Training and Evaluation

The dataset of 892 samples was divided randomly as 50% for training and 50% for testing. The same training and test sets were used in all experiments. The training set was normalized to unit norm. This normalization parameter was transferred to the test set and is the only data preprocessing step applied to the dataset. No alignment technique in time (frames) or space (sensor amplitude) was applied since the objective of our experiments is to let the CNN learn from the spatio-temporal raw sensor data. For training, the multiclass log loss was used as an objective function. The Adam optimizer [27] was used with backpropagation to update the network weights. A small batch size of 4 was used to calculate the gradient of the loss function per epoch. Our models were trained from scratch and the weights of the network were initialized following the Glorot uniform initialization technique [28]. An early stopping rule was applied: The model stopped the training procedure if the validation error did not improve after 20 epochs. The metrics used for evaluation of the model's performance were accuracy, precision, recall, and *F*-score.

VII. RESULTS AND ANALYSIS

A. Cross Validation and Feature Extraction for CNNs

The experiments presented in this section are cross validated by using a tenfold cross validation technique due to its good statistical performance [29]. All the results are presented with a 95% statistically confident interval. For the experiments that use CNNs, the dataset features created by the network are extracted before the classification layer (softmax) to allow a faster and more efficient calculation of the cross-validated results by using a linear classifier. The extracted features are then trained and evaluated with a linear SVM model to obtain cross-validated scores.

B. Classification Performance of Spatio-Temporal RSMs

The CNN model presented in Section VI-B was used to test the effect on the performance of varying the number of consecutive frames per sample of the dataset by using the spatio-temporal RSMs. The chosen ten sequence sets cover the entire available frame range from 1 to 280. Table II shows the performance metrics results for the frame sequence sets, top performance indicated in bold. It is observed that the best classification performance score is obtained for a frame set of 30 with an *F*-score of 97.88% ± 1.70%, which corresponds to an experiment duration of 0.58 s out of 5.6 s for the longest sequence (280 frames). Classification performance declines for sequences longer than 30 frames and is lowest in the extreme case of a single frame, the latter observation attributed to insufficient data. Thus, only the first 30 consecutive frames were considered in the reported experiments since they also capture a single footfall, these results are the best obtained for the minimum amount of data that needs to be processed.

Models of 30 consecutive frames appear to be optimal since they involve an acceptably small amount of data (about 11% of the executed gait pattern), while yielding the best *F*-score. Increasing the number of frames above 30 decreases the *F*-score slightly and is sustained up to the highest number of frames.

C. Top Performing CNN Model

This section presents the top performing CNN model based on the architecture presented in Section VI-B for the UoM-gait-13 dataset considering 30 consecutive frames per sample as indicated in Section VII to fine-tune our CNN model. The learning rate of the Adam optimizer was set to 0.0001 and the other optimizer parameters were set at the recommended values provided in [27]. The confusion matrix of the top performing CNN model is shown in Fig. 6(a). The training and validation loss per epoch can be observed in Fig. 6(b). For this CNN model, we obtained an accuracy of 97.88 ± 1.70%, precision of 98.09 ± 1.50% recall of 97.88 ± 1.70% and *F*-score of 97.88 ± 1.70%, which were the best classification performance scores observed overall using the UoM-Gait-13 dataset.

D. Comparison of Spatio-Temporal RSMs and Tomography Reconstructed Spatio-Temporal Features

In this section, we compare classification performance based on automatic feature extraction from raw sensor data with classification performance based on features from reconstructed tomography images. This is achieved by comparing the best performing deep learning CNN model with RSM input as

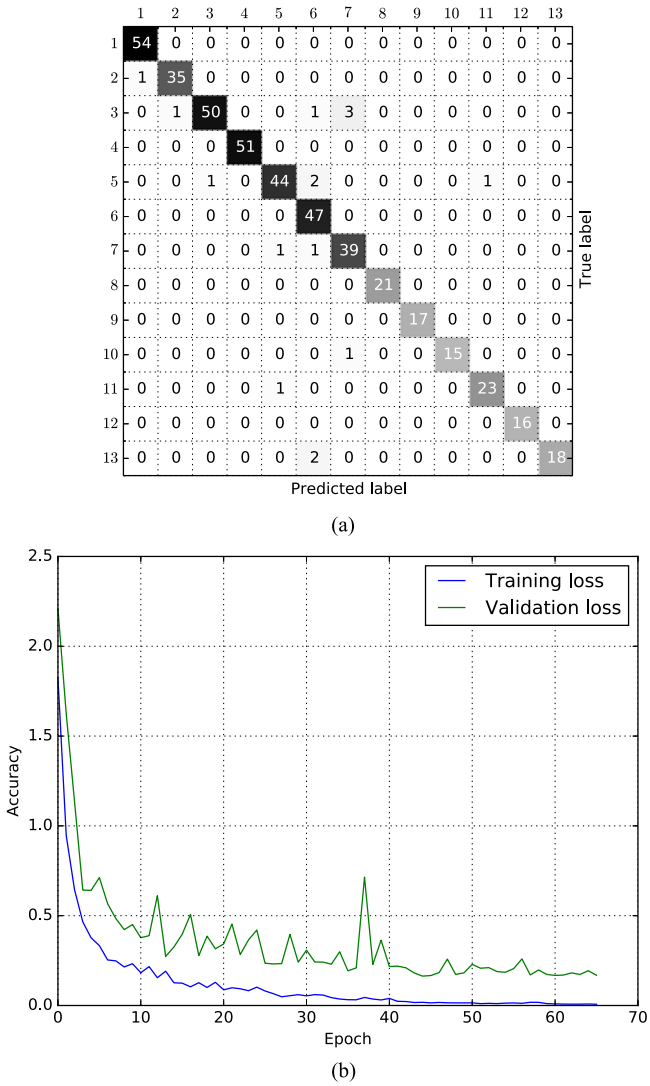

Fig. 6. Performance characteristics of the fine-tuned CNN model (a) confusion matrix showing the classification performance of the learning model: predicted class versus actual class (b) training and validation loss curve of the model as a function of epoch.

presented in Section VII-C against shallow learning models using the spatial image reconstructions as presented in Section V-B. For the latter, each of the 200 × 100 pixel matrices of gait experiments samples were considered as a feature vector of 20,000-pixel features and used in a linear SVM, nonlinear SVM with Radial Basis Function (RBF) kernel and Random Forest models for performance comparison. The model hyperparameters for the SVM model with RBF kernel are penalty parameter of C:1 and a gamma parameter of 0.00005. For the Random Forest model, 100 decision tree estimators were used. For both models, a grid parameter search technique was used for hyperparameter selection.

Table III shows the performance comparison of the models, top performance is indicated in bold numbers. The CNN model using the spatio-temporal RSMs outperforms by a wide margin the models that use the spatially reconstructed tomography sensor data, as well as drastically the nonlinear SVM that uses the same spatio-temporal RSMs. The linear SVM model performed best for the spatial reconstruction approach, followed by the use of the CNN model on the reconstructed spatial data, however, these results did not approach the range achieved by the CNN approach. This comparison justifies the effectiveness of our CNN model and raw sensor data representation as RSMs.

E. Filter Maps Visualization

By using the RSMs as input to the trained CNN model presented in Section VII-C, we extracted the output volumes of the ReLU activations of the second conv layer to observe the feature maps learned by the model as presented in Fig. 7. The frames of the normal walk experiment as presented in Fig. 3 were selected as well for feature maps learned by the model. This allows a pixel-by-pixel comparison between raw data (see Fig. 3) and feature maps (see Fig. 7) due to the stride (subsample) of 1 used in the conv layers of the CNNs. Fig. 7 illustrates the automatic discrimination of features by the network resulting in a fairly limited set of active measured sensor amplitudes. While the RSMs sequence from frame 1 to frame 30 in Fig. 3 is consistent with persisting clusters of low or high measured signal amplitudes (light and dark), the filter maps in Fig. 7 appear to have less activated sensor amplitudes compared with Fig. 3. This shows the ability of the CNN to discriminate unnecessary sensor data and to work with an optimal combination of sensor amplitudes in the processing steps between layers of the CNN.

F. Model Training Time Execution Comparison

1) CNN Models: All the models were trained on a Titan X GPU, that accelerates model training by using the cuDNN library [30] with optimized GPU-based neural network training routines. Table II shows the classification performance of the CNN models according to the number of frames considered. The training time of the models is in the range of 34.66–92.12 s, resulting from the number of training epochs and the execution time per epoch. The latter shows the amount of model training time required for a single pass of the training set in the CNN. The average execution time per epoch is 0.978 s $\pm$ 0.36 s. The observed range of performance justifies our optimal model choice of 30 frames, reducing the execution time per epoch to 0.78 s from 1.92 s for the 280 frames model.

2) Shallow Models and Optimal CNN Model: The models in Table III were trained at the CPU level, except for the optimal spatio-temporal RSM model that was trained at the GPU level. Here longer training times are observed for the SVM models with the reconstructed image features. This effect is caused mainly due to the high dimensionality of the feature vector since the execution time is strongly dependent on the number of features and increases in parallel with the memory consumption of each training sample. For the RSM experiments, the feature vector is 110 pixels long, Increasing to 20,000 pixels for the reconstruction models. Furthermore, the execution time increases together with the model complexity. This is illustrated by the training time comparison for spatial reconstruction feature between a linear SVM model (219.21 s) and an SVM model with RBF kernel (249.23 s); the latter requiring more time to train due to its nonlinearity and increased complexity. We also performed

TABLE III
RAW SENSOR FEATURES AND SPATIAL RECONSTRUCTED FEATURES MODEL COMPARISON

Experiment Description	Accuracy	Precison	Recall	*F*-score	Training Time (s)
Spatio-temporal RSMs (30 frames)	**97.88 ± 1.70%**	**98.09 ± 1.50%**	**97.88 ± 1.70%**	**97.88 ± 1.70%**	**51.68**
Spatio-temporal raw data (30 frames) (RBF Kernel SVM)	48.32 ± 2.57%	41.49 ± 2.73%	48.32 ± 2.57%	41.18 ± 2.42%	250.70
Spatial reconstruction (CNN model)	70.77 ± 3.44%	72.96 ± 4.22%	70.77 ± 3.44%	68.67 ± 3.43%	35.90
Spatial reconstruction (Linear SVM)	78.11 ± 2.72%	78.58 ± 2.89%	78.11 ± 2.75%	77.36 ± 2.80%	219.21
Spatial reconstruction (RBF Kernel SVM)	68.69 ± 2.20%	65.93 ± 2.27%	68.69 ± 2.20%	65.86 ± 2.19%	249.23
Spatial reconstruction (Random Forest)	75.89 ± 2.49%	75.6 ± 2.76%	75.90 ± 2.48%	74.41 ± 2.50%	10.67

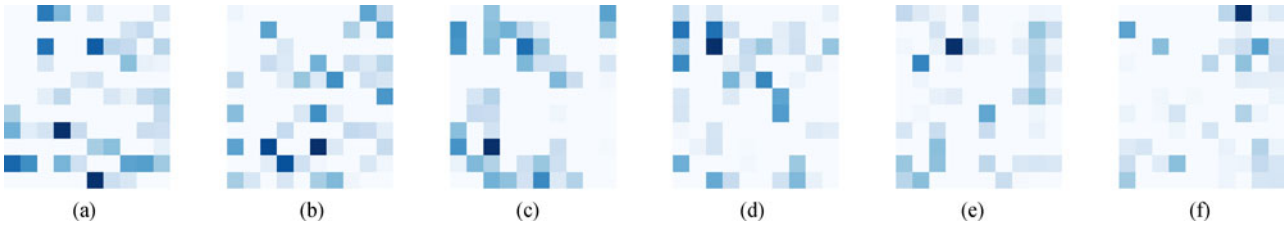

Fig. 7. Activation maps of a set of frames from one sample of the normal walk experiment. (a) Frame 1. (b) Frame 10. (c) Frame 15. (d) Frame 20. (e) Frame 25. (f) Frame 30.

an experiment with a fast implementation of the Random Forest algorithm [31], as manifesting the optimal overall training time of 10.67 s.

VIII. DISCUSSION AND CONCLUSION

Compared to tomography sensing, resulting in the reconstruction of images, which then need interpretation, we render the measured data as a combination of different sensor amplitudes over time, which the CNN has to identify and learn automatically, this capability is shown when comparing the feature maps presented in Fig. 7 and the RSM representation presented in Fig. 3. The spatio-temporal RSMs are equivalent to a Radon transform [18], as they contain the same measurement information, but the ordering of the data (e.g., as angular projections in a sinogram, which contains the spatial information) become irrelevant for the CNN approach presented here. Nevertheless, the latter is successful in accurate spatio-temporal gait data classification, from raw tomography sensor data, without the need to reconstruct images. Because of the sensor head geometry (the sampling of the forward Radon transform), designed with tomography in mind, ensures adequate spatial sampling to capture the spatial distribution with the needed resolution both in space and time. While the achievable spatial sensitivity is determined by the sensor head geometry according to general tomography theory, the time resolution refers to the time of the gait experiments. Both effects are incorporated in the spatial average signals as shown in Fig. 2.

The justification for demonstrating the methodology with data from a floor sensing system is multifaceted. The complexity of the problem addressed here lies not only in identifying the types of footstep patterns but also in the necessity to cope with the footstep spatial and temporal variances in the data captured from a real-world gait recording system. The variances in space refer to the fact that in the gait experiments the user can step anywhere on the floor sensor surface, at any given time while walking, which activates sensors with different response characteristics. This is invariant with the subject of tomography sensing and, therefore, is not limited to the particulars of the used demonstrator. The general issue of classifications based on spatio-temporal raw data versus features derived from image reconstructions are similarly invariant across tomography modalities, as justified in Section I.

In conclusion, we have introduced a CNN model architecture that automatically learns a robust set of feature representations from raw spatio-temporal tomography sensor data and eliminates the tomography image reconstruction step. In a case study with experimental data from a demonstrator floor pressure sensing system, the approach proved to be highly successful as evidenced by the *F*-score of 97.88 ± 1.70% achieved with the optimized CNN model presented in Section VII-C. To the best of our knowledge, a problem and a solution of this nature have not been reported in the literature before. This approach is portable across a range of industrial applications requiring tomography sensing.

Future goals, advised by the results presented in this paper, will be to construct models that learn spatio-temporal representations from raw sensor data as the monitoring task evolves in time. Such always-on monitoring systems may allow the detection of the onset of unforeseen changes in an industrial process, such as unintended modification of process components, vessels, flows, etc.

ACKNOWLEDGMENT

The authors express their gratitude to David H. Foster for useful discussions. They acknowledge NVIDIA for the donation of the GPU used to perform some of the experiments of this research. The author O. Costilla-Reyes would like to acknowledge CONACyT (Mexico) for a studentship.

REFERENCES

[1] T. York, H. McCann, and K. B. Ozanyan, "Agile sensing systems for tomography," *IEEE Sensors J.*, vol. 12, no. 11, pp. 3086–3105, Dec. 2011.

[2] K. B. Ozanyan, "Tomography defined as sensor fusion," in *Proc. IEEE SENSORS*, Nov. 2015, pp. 1–4.

[3] K. Choi, J. Wang, L. Zhu, T.-S. Suh, S. Boyd, and L. Xing, "Compressed sensing based cone-beam computed tomography reconstruction with a first-order method," *Med. Phys.*, vol. 37, no. 9, pp. 5113–5125, 2010.

[4] F. Barthel, M. Bieberle, D. Hoppe, M. Banowski, and U. Hampel, "Velocity measurement for two-phase flows based on ultrafast x-ray tomography," *Flow Meas. Instrum.*, vol. 46, pp. 196–203, 2015.

[5] S. Rabha, M. Schubert, and U. Hampel, "Regime transition in viscous and pseudo viscous systems: A comparative study," *Amer. Instit. Chem. Eng. J.*, vol. 60, no. 8, pp. 3079–3090, 2014.

[6] J. Mohamad-Saleh and B. Hoyle, "Determination of multi-component flow process parameters based on electrical capacitance tomography data using artificial neural networks," *Meas. Sci. Technol.*, vol. 13, no. 12, pp. 1815–1821, 2002.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] J. A. Cantoral-Ceballos *et al.*, "Intelligent carpet system, based on photonic guided-path tomography, for gait and balance monitoring in home environments," *IEEE Sensors J.*, vol. 15, no. 1, pp. 279–289, Jan. 2015.

[9] S. K. Saripalle, G. C. Paiva, T. C. Cliett, R. R. Derakhshani, G. W. King, and C. T. Lovelace, "Classification of body movements based on posturographic data," *Human Movement Sci.*, vol. 33, pp. 238–250, 2014.

[10] J. M. Hausdorff, A. Schweiger, T. Herman, G. Yogev-Seligmann, and N. Giladi, "Dual-task decrements in gait: Contributing factors among healthy older adults," *J. Gerontology Ser. A: Biol. Sci. Med. Sci.*, vol. 63, no. 12, pp. 1335–1343, 2008.

[11] M. W. Whittle, *Gait Analysis: An Introduction*. London, U.K.: Butterworth, 2014.

[12] S. Gay, J. Santos-Victor, and A. Ijspeert, "Learning robot gait stability using neural networks as sensory feedback function for central pattern generators," in *Proc. 2013 IEEE/RSJ Int. Conf., Intell. Robots Syst.*, 2013, pp. 194–201.

[13] I. Birch, T. Birch, and D. Bray, "The identification of emotions from gait," *Science Justice*, vol. 56, pp. 351–356, 2016.

[14] A. Muro-de-la Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, 2014.

[15] M. Ziefle, S. Himmel, and W. Wilkowska, "When your living space knows what you do: Acceptance of medical home monitoring by different technologies," in *Proc. Symp. Austrian Human Comput. Interact. Usability Eng. Group*, 2011, pp. 607–624.

[16] R. Headon and R. Curwen, "Recognizing movements from the ground reaction force," in *Proc. 2001 Workshop Perceptive User Interfaces*, 2001, pp. 1–8

[17] L. Landweber, "An iteration formula for fredholm integral equations of the first kind," *Amer. J. Math.*, vol. 73, no. 3, pp. 615–624, 1951.

[18] K. B. Ozanyan, S. G. Castillo, and F. P. Ortiz, "Guided-path tomography sensors for nonplanar mapping," *IEEE Sensors J.*, vol. 5, no. 2, pp. 167–174, Apr. 2005.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[20] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[22] O. Costilla-Reyes, P. Scully, and K. B. Ozanyan, "Temporal pattern recognition in gait activities recorded with a footprint imaging sensor system," *IEEE Sensors J.*, vol. 16, no. 24, pp. 8815–8822, Dec. 2016.

[23] R. W. Daniels, *Approximation Methods for Electronic Filter Design: With Applications to Passive, Active, and Digital Networks*. New York, NY, USA: McGraw-Hill, 1974.

[24] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," arXiv: 1605.02688, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[26] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[27] D. Kingma and J. Ba, "ADAM: A method for stochastic optimization," arXiv:1412.6980, 2014.

[28] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[29] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Int. Joint Conf. Artif. Intell.*, 1995, vol. 2, pp. 1137–1143.

[30] S. Chetlur *et al.*, "CUDNN: Efficient primitives for deep learning," arXiv:1410.0759, 2014.

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

Omar Costilla-Reyes received the M.Sc. degree in electrical engineering from the University of North Texas, Denton, Texas, USA, in 2014. He is currently working toward the Ph.D. degree in electrical and electronics engineering at the School of Electrical and Electronics Engineering, University of Manchester, Manchester, U.K.

During his master studies, he was a Research Assistant in projects with funding from the National Science Foundation and National Aeronautics and Space Administration. His M.Sc. dissertation was on dynamic indoor positioning systems using wireless sensor networks. He is currently a Research Associate with the University of Manchester. He has published papers on machine learning for security and healthcare. His research interest focuses on machine learning for security and healthcare with sensors systems.

Omar Costilla-Reyes received the Best Student Paper Award in Optical Sensing applications at the 2015 IEEE Sensors Conference. His Journal paper entitled "Temporal Pattern Recognition in Gait Activities Recorded With a Footprint Imaging Sensor System" was one of the 25 most downloaded IEEE Sensors Journal papers from January to May 2017. He received scholarships and awards for academic achievement including an academic scholarship for his master's and doctorate studies from the Mexican Science council (CONACyT).

Patricia Scully received the Ph.D. degree in engineering from the University of Liverpool, Liverpool, U.K., in 1992.

She was a Reader with Liverpool John Moores University, Liverpool, U.K., in 2000, before joining the University of Manchester as a Senior Lecturer/Associate Professor of sensor instrumentation in 2002. She is experienced in leading industrial and research council/government funded research projects at national and international levels, and her research interests include sensors and monitoring for industrial processes, including optical fibre technology and photonic materials for sensors and devices, ranging from functional chemically sensitive optical coatings, to laser inscribed photonic and conducting structures in transparent materials that affect the properties of light.

Krikor B. Ozanyan (SM'95) received the M.Sc. degree in engineering physics (semiconductors) and the Ph.D. degree in solid-state physics from the University of Sofia in 1980 and 1989, respectively.

He is currently the Director of Research in the School of Electrical and Electronics Engineering, University of Manchester, Manchester, U.K. He is a Fellow of the Institute of Engineering and Technology, Stevenage, U.K., and the Institute of Physics, Marylebone, U.K. He has more than 300 publications in the areas of devices, materials and systems for sensing and imaging.

Dr. Ozanyan was a Distinguished Lecturer of the IEEE Sensors Council in 2009 and 2010, and the Guest Editor for the 10th Anniversary Issue of the *IEEE Sensors Journal* in 2010, as well as the Special Issues on "Sensors for Industrial Process Tomography" in 2005 and "THz Sensing: Materials, Devices and Systems" in 2012. He is currently in his second term as an Editor-in-Chief of the *IEEE Sensors Journal* and is a General Co-Chair of the IEEE SENSORS 2017 Conference.